Design Considerations for an Indoor Location Service Using 802.11 Wireless Signal Strength

by

David M Lambeth

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Masters of Engineering in Aeronautics and Astronautics
[Master of Engineering]
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2009

Author .....
Department of Aeronautics and Astronautics
January 30, 2009

Certified by .....
Seth Teller
Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by .....
Prof. David L. Darmofal
Associate Department Head
Chair, Committee on Graduate Students

Design Considerations for an Indoor Location Service Using 802.11 Wireless Signal Strength

by

David M Lambeth

Submitted to the Department of Aeronautics and Astronautics
on January 30, 2009, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Aeronautics and Astronautics

Abstract

This thesis compares approaches to the problem of discovering a mobile user's location indoors. The particular challenges of location discovery using 802.11 (Wi-Fi) signals and "organically collected" (i.e. user-generated) received signal strength maps are discussed. Several existing and novel localizer algorithms are compared on a database of organically collected data. Features of local Wi-Fi "signatures" which are relevant to location discovery are identified and applied to algorithm design considerations. Future directions for algorithm refinement are discussed.

Thesis Supervisor: Seth Teller
Title: Professor of Computer Science and Engineering

Acknowledgments

The author would like to recognize the financial and technical support of Nokia Research Center Cambridge and MIT's Computer Science and Artificial Intelligence Laboratory. The author would also like to thank my advisor and project leader Dr. Seth Teller; Dr. David Clark; Dorothy Curtis, Ben Charrow, and RJ Ryan for building our experimental application; Yoni Battat for developing our indoor mapping tools; and Jonathan Ledlie and Jamie Hicks from Nokia Research Center Cambridge for their advice.

Contents

List of Figures

Chapter 1

Introduction to the Indoor Localization Problem

The increasing computing capabilities of mobile devices such as cell phones and personal digital assistants has fed interest in creating applications which take advantage of their mobility. Mapping services are already available for mobile devices which give directions and even guided tours to walking users using GPS [8]. There is interest in extending such services indoors as well. Location-specific web searching has been available for several years [9]. Sense Networks aggregates anonymized user motions gathered from GPS-enabled mobile devices to track economic activity, locate spots of interest, and identify "tribes" of users with similar behavior and tastes [32]. Other location-enabled services are available in prototype or beta form. The Locale application for Google Android mobile phones allows users to specify how settings should change based on location, for instance automatically silencing ring tones in a lecture hall [10]. The MIT iFind project combines an instant messenger client with the ability to see your friends' locations on a map and calculate convenient meeting places [14].

All of these services require an estimate of a user's location in the world, and they typically rely on a small, inexpensive GPS receiver built into the mobile device. While GPS works reasonably well in open outdoor environments, its signals cannot reach indoors or into "urban canyons" in dense cities where little of the sky is visible.

Furthermore, the location estimate returned by a GPS receiver is in globally georeferenced coordinates for which few indoor maps exist. In order to provide location-based services indoors where users spend much of their time, an accurate way of determining user location indoors is required.

Early indoor location discovery systems used dedicated fixed hardware such as ultrasound transponders [35] or infrared beacons [40] to replace the satellites which provide reference positions in GPS. While effective, these systems suffered from a significant up-front cost to purchase, install, and configure the fixed hardware anywhere one wanted to use the system. Since the late 1990s, the proliferation of wireless Internet or Wi-Fi access points using the 802.11 family of standards has provided a set of indoor radio frequency beacons which can be used by location discovery systems for free. These have been used in both experimental and commercial location discovery outdoors [27, 16, 6] and indoors [3, 22, 49, 34, 15, 17].

The sheer variety of approaches to indoor location using Wi-Fi makes it difficult to compare different designs. Even when algorithms for localization are presented in published literature, they may contain undocumented but important refinements which make it difficult to replicate their performance in a clean-slate recreation [11]. Comparisons using common data sets, common methodology, or even common performance metrics are rare. This thesis aims to identify features of Wi-Fi networks important to evaluating localizers, to compare a family of simple designs on the same data set, and to identify refinements which can be pursued by future researchers hoping to build an indoor location discovery service.

Chapter 2

Signature-Based Methods

2.1 Indoor signal propagation and radio signatures

The indoor location problem poses many challenges beyond establishing a set of beacons. Radio propagation inside buildings is complex. Relatively few lines of sight exist spanning more than one room. Radio signals are attenuated by traveling through walls, furniture, and even people. Walls, floors, and ceilings cause signals to partially reflect, allowing them to reach a receiver by multiple paths.

Early attempts tried to correct for these effects using signal propagation models and perform trilateration like GPS [3]. An alternate approach is to examine ping-response times using multiple back-and-forth echo messages [53]. Both of these methods require a detailed knowledge of access point locations. Furthermore, simple signal propagation models are only accurate over line of sight paths [25, 2]. To account for multipath effects, indoor signal propagation models require detailed knowledge of building and network geometry, and still tend to underperform empirical surveys [18]. Although research continues on ever more refined models, such as particle filters [43], most indoor location discovery systems have abandoned this approach. The facts that the indoor radio environment is affected by unpredictable factors such as furniture and people [23], and that received signal strength can vary by 10 dB or more on a scale of tens of centimeters [49], mean that it is easier to ignore the signal path and focus on the result at the receiver.

Wi-Fi access points periodically broadcast a "beacon frame" which announces their identity, including MAC address and network name, to any wireless devices within receiving range. Standard wireless card interfaces have the capability to passively "scan" for such beacon frames or actively request them from all nearby access points. The set of access points whose beacon frames are visible to a receiver, combined with the received signal strength for each transmitter, form a wireless "signature" (also known as a wireless "fingerprint" [19]). Because a wireless signature is a sampling of the radio environment for 802.11 signals, it varies from one place to another indoors. In an ideal, noiseless environment with sufficient access point density, each signature would uniquely identify a point in space regardless of the complexity of signal propagation. (The messy reality is discussed in section 5.)

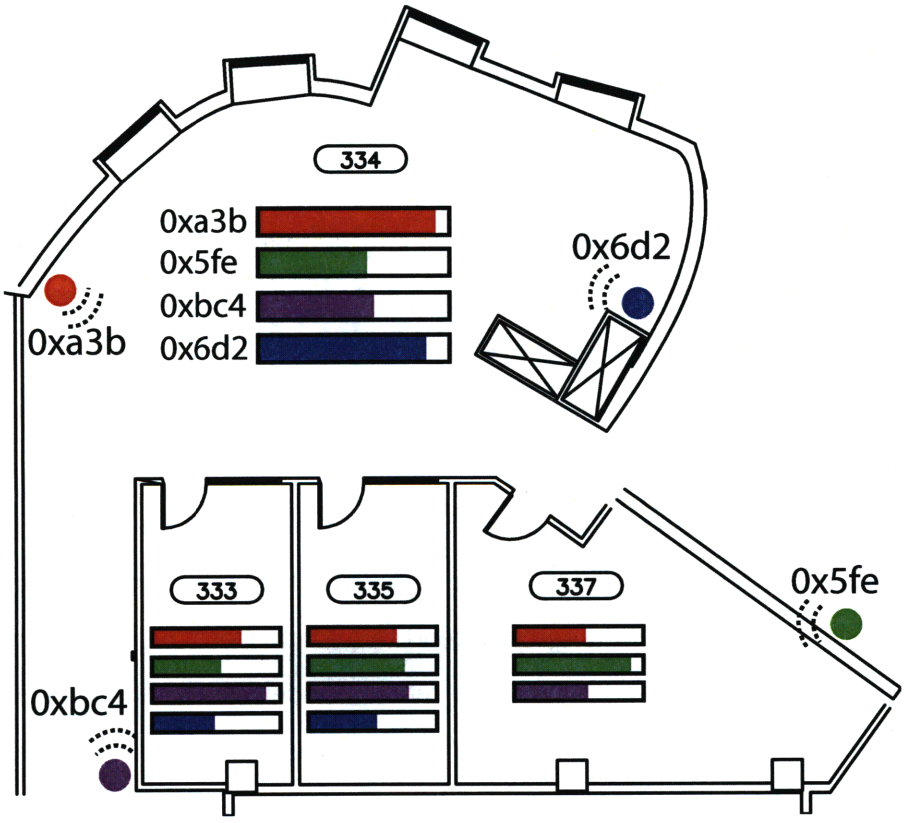

Figure 2-1: An example of idealized RF signatures. The RSS of an access point in a location (or whether the AP is even visible there) depends on complicated indoor signal propagation from the AP to that location.

There are a wide variety of approaches to localization using radio signatures, and there have been attempts to create a formal taxonomy of the approaches [19]. Without going into the details of the various taxonomic criteria, the algorithms discussed in this thesis are of a type intended for location-enabled services: empirically derived radio maps with private client-side location discovery and potentially unlimited scalability.

2.2 Survey based approaches

The basic concept of survey-based localization with Wi-Fi is straightforward. A set of reference points is selected to cover the area throughout which one wants to implement indoor location discovery. These points could follow a regular grid [3, 49], they could be placed one to a room [26], or they could even be strung along linear paths through corridors and rooms [15]. Linear features are useful in places where the user is restricted to only a few possible paths, such as in corridors. As the receiver moves through the radio environment, its readings can be matched against stored paths, provided the speeds and sampling rates are roughly similar [29]. Grid surveys yield the smallest absolute position errors and even the ability to coarsely extract user orientation [3, 23]. Room-level surveys have lower absolute accuracy but tend to be sufficient for human users, since in most rooms the user can see the whole room from any point inside it [36]. A robot would most likely use Wi-Fi for global localization and rely on other sensors for finer navigation. Room-level accuracy should be sufficient for that purpose as well. In theory one could provide a finer-grain location result by combining a room-level survey with signal propagation models, but this approach does not approach the performance of a denser grid survey [20].

2.2.1 The survey process

In the initial "training" phase of the system, a surveyor with some simple training in how to collect signatures visits each reference point or region in turn with a wireless device. The device is configured to repeatedly scan the local wireless network, producing a list of the access points visible and some samples of the received signal

strength for each. These measurements are combined to form the wireless signature of that location. Attempts have been made to determine the minimum number of scans required at each location to build a useful signature. Answers tend to converge on approximately 100 scans, requiring one to two minutes of data collection depending on the wireless device [26, 49], though in one case as little as 20 seconds of data was collected per location with only a 12% performance penalty [20].

In the second, "use" phase, wireless devices carried by users scan the local network. A localizer algorithm matches these user scans against the stored signatures from the survey to determine which one is the closest match. The best-matching signature indicates the user's most likely location. The algorithm could be on a central server (for faster computation) or on the client's device (for greater privacy), as long as it has access to current scans and a corpus of signatures. Additional steps in the localization algorithm might average several of the best-matching survey locations for more accuracy [47] or use a Hidden Markov Model to improve tracking of a mobile user [26].

2.2.2 Drawbacks to surveys

The amount of preparation required for a survey can be significant. The surveyor typically requires maps of the area (such as a set of floor plans) with markings at the locations to be surveyed [26]. Access point locations need not be known, an advantage over signal propagation methods, but the local network must meet some basic criteria. The number of access points required for accurate localization is unclear, but is more than the minimum required for network connectivity. Whether a signature could match multiple surveyed locations naturally depends on the variation in signatures between those locations. The finer the survey, the more access points will be required to resolve ambiguities. Our tests were conducted in a building with an unusually high density of more than 200 access points over ten floors. Surveyor training can also be a significant challenge, especially for grids or linear features. Surveying a linear feature requires the surveyor to carefully note their path and walk at an unnaturally constant, slow speed so that scans can be matched to locations by linear interpolation [15].

Surveys require the cooperation of the owners and occupants of the area where the system will be deployed. By surveying the existing radio signatures, these systems avoid the need to install any potentially burdensome permanent infrastructure in the operating environment. However, they do require that the surveyors be given temporary access to the whole area. Getting permission to access places like private offices can be difficult and complicates the survey process [26]. In some cases, getting physical access may be impossible, leaving gaps where location discovery will be inaccurate until survey data can be collected to fill the gaps.

The performance of a localizer based on Wi-Fi surveying depends heavily on the nature of the original survey as well as the details of the matching algorithm. The latter effects will be investigated in this thesis. Grid surveys typically report an error distance of half the distance between adjacent survey points, or larger [49, 3]. Surveys which characterize each room with one signature report successfully identifying which room a user is in up to 95% of the time with error distances of 5 meters or less [26]. Ekahau's system, which surveys along linear paths, reports an accuracy of 1 to 3 meters for stationary users [15]. Reported error distances should be taken with a grain of salt, since they can be difficult to calculate if the localizer's estimate is returned as a region such as a room or a linear segment. Outside the surveyed area, localizer performance degrades rapidly [38].

The time required to survey even a single building can be prohibitive. The fastest one could reasonably survey is one minute per room [26]. The office building at MIT in which we deployed our test system has over 1400 unique room-sized locations. Surveying it fully would require a minimum of 24 person-hours. In order to have a reference comparison for our own system, we had Ekahau perform a survey to deploy their system over most of the building. This undertaking required two of their employees to visit our building in person and spend three working days slowly pacing rooms and corridors to build their signature database. Many rooms in the intended deployment area could not be surveyed because their occupants could not be located in order to request access. Deploying a system over an entire campus with many buildings, let alone over a city or continent, would require an immense amount of

dedicated effort. Linking existing surveys from different sources to expand coverage is hampered by differences in survey methods and data representation.

Even if the required up-front effort is expended to establish a survey, any changes that affect the radio environment will negatively impact the system's accuracy. Access points may be moved or replaced over time, and the maintainers of the location discovery system might not be notified. Alterations to the arrangement of walls or even furniture can cause local changes in the wireless signatures. The only way to restore performance is to re-perform the survey in areas affected by the changes, assuming one can pinpoint such areas. Staleness of data affects every survey-based system to some degree, and many commercial providers of such systems plan for regular re-surveying as part of every deployment [16, 15].

2.3 Organic data collection

At MIT's Computer Science and Artificial Intelligence Laboratory we have been developing a location discovery service which circumvents the difficulties of survey-based services. We combine the training and use phases by enlisting the system's users to gradually build the database of wireless signatures. This we call "organic data collection." There are precedents for user contribution in outdoor location discovery. The practice of "war driving," or driving streets in search of open Wi-Fi access, has yielded online databases containing location information on over 14 million wireless networks [44]. Intel's PlaceLab project used these databases and known GSM cell tower locations to provide an experimental outdoor location discovery service [27]. They also examined expanding their beacon database by incorporating additional beacons found by users [28]. In contrast to PlaceLab, our organic data collection is focused on indoor spaces and on signatures rather than beacon locations for the reasons described in section 2.1.

The key to organic collection is to minimize the effort required of users who wish to contribute. A single click on a digital map is enough to associate a few contemporaneously gathered scans with a location. We also allow users to specify a

time interval up to an hour into the past and/or future during which they were, or will be, at the indicated location. This yields from dozens to thousands of scans from which to build a signature [39]. A large corpus of signatures which would require hours of dedicated attention from a team of surveyors can be organically collected with only occasional and momentary attention from users because the users do not have to modify their daily movements to contribute.

The degree to which users contribute to an organic database varies significantly, as seen in many other collaborative efforts such as Wikipedia [42]. A few motivated users will perform small-scale surveys of significant areas. Less enthusiastic users will contribute data for a few places they frequent, such as their home or office. The remainder of the users contribute few or no scans, but they can use signatures contributed by other users who came before them to localize themselves. The breakdown between high, medium, and low contributors for a small experiment in organic data collection can be seen in figure 4-1.

2.4 Advantages and limitations to organic collection

The advantages of organic data collection stem from its flexibility. The available coverage area for location discovery expands as the user base of the service expands. The spaces most often visited by a user can be covered by them with only a small effort and will thereafter be available to any user. Questions of obtaining access to places are irrelevant. If changes to the network or the physical environment degrade performance (and the area affected by the change can be determined), users can be prompted to contribute fresh data to restore performance. There is no need to wait for surveyors to return to the area.

Organic data collection can neatly sidestep access issues which hamper surveys. The PlaceLab group analyzed the diaries of volunteers who were asked to record their locations over a one-month period. They found that only a few frequently-visited

places accounted for the majority of volunteers' time [12]. These are often places that the user frequents but other people would visit only occasionally, such as the user's home or office. Access to these private or semi-private places can be difficult for a stranger conducting a survey. If users find location-enabled applications sufficiently compelling, though, they are likely to contribute signatures for these few locations within the first few days of using the service, thus circumventing access problems. Because these signatures can be shared, all users can have accurate location even in non-public places. This process can result in more complete coverage than a dedicated survey would.

A less obvious benefit to organic collection arises from the way users contribute data as they go about their daily business. This means that in organically collected data, the scans comprising the signature for a room would be biased toward those parts of the room which users visit most frequently. For instance, a dedicated surveyor might try to characterize the radio environment of an office by walking around the walls and across the center to get an even pattern of scanned points [26]. The occupant of the office, in order to organically contribute data, may just set their wireless device down at their desk. If most visitors to the office also sit at or near that desk, which signature is more likely to accurately match the scans acquired by their device? It is difficult to get a quantitative answer to this question, but if most users frequent the same parts of a room, then there is a tendency for organically collected signatures to match the room's typical use patterns. On a larger scale, organic data collection covers only those spaces that are actually used. This potentially makes it more efficient than deliberate surveying in terms of effort and data storage.

That said, there are advantages to having trained, dedicated surveying staff. Each signature has to be associated with an area which an untrained user can understand. Areas consisting of a single room or portions of a larger area the size of a typical room (typically 10 to 25 m^2) form the most natural partitioning for human users [20]. We trade the potentially greater accuracy of a grid survey for ease of organic contribution.

While we run our location discovery algorithm only on the user's device in order to keep the result private, an organic system still needs occasional connectivity

to a central server. The client will need to upload contributions and download updated signatures that incorporate other users' recent contributions. The user will also need to download new signatures if he or she moves to a new area (see section 3.3). Passively using the service does not reveal a user's location, but contributing data potentially could [39]. Users therefore face a trade-off between privacy and contribution. Privacy is unlikely to be a significant concern, though, since contributions can be made anonymously without compromising system performance and some users may welcome recognition of their contributions.

Organic data collection is subject to errors in the data contributed by users. They may simply be mistaken about their location (or the period of time they will be there), or they may be attempting to maliciously mislead the system. We have not yet addressed these potential issues, but some approaches will be discussed in section 7.2.

Even given room-level accuracy and benign, conscientious users, there are statistical properties of organically collected data which make it more challenging to use in a localizer algorithm than data from a planned survey. Organically collected data is not expected to be spread evenly over an area like surveyed data, but to clump according to the density of users in the area and their savviness. For some locations users contribute only a few dozen scans while for others they contribute tens of thousands [39]. A localizer algorithm designed for organic data must compensate for these sampling disparities, something which survey-based localizers can ignore. Section 3.2 will address this issue.

Chapter 3

Representation

3.1 Data structures

All Wi-Fi enabled devices which comply with the 802.11 standards have a method for scanning the local network. When the method in the wireless driver is invoked, the wireless card searches the 12 (or 14) Wi-Fi channels for the beacon frames transmitted by nearby access points. The time required for a scan of the wireless network differs between drivers. For the standard wireless driver on device we used in our experiments, the Nokia N810 tablet computer, the mean time between scans is 2 seconds. Providing accurate location discovery services to a mobile user requires that fresh scans be taken as frequently as possible, while still finding most of the visible access points. Recent work towards providing Wi-Fi connectivity in moving vehicles has produced methods which can reduce the time required to complete a passive network scan by 40% by prioritizing the most frequently used channels [7]. During the scan, the wireless driver records each unique transmitting MAC address seen and a measurement of received signal strength. Different drivers return measurements with different ranges and scales, but in general they seem to be linearly related to dBm intensity of the measured signal [26]. Calibration of different drivers to a common scale is discussed in section 7.3.

We call each pair of a MAC address and a RSS value a reading. A reading is a single measurement of a single access point at a single time. A collection of readings

for all APs visible to the wireless card at a single time, which is what the driver returns, is known as a scan. When all the scans taken in a single place are collected together, the result is the wireless signature of that place, covering many different times and access points. The collection of all readings for a single access point in a single location is also known as the signature of that access point. Which meaning of signature is intended will be clear from context. When a user indicates his or her location for some time period, the scans collected by his or her device during that period and sent to the signature database are called a "bind." A bind is identical in structure to a location signature, except that it comes from a single user over a single interval of time. A signature for location "A" is the composition of all binds labeled with location "A" from all users.

3.2 Compression of data

Recording every reading taken in a location separately can result in a signature of unwieldy size. Because the Nokia N810 tablets we used return a new scan every 2 seconds, a stationary tablet collecting data for just a few hours would obtain a bind containing thousands of scans. Much of this data is redundant, yet the size of it would slow down the localizer algorithm. Also, sharing all of this data with other users would make poor use of the limited bandwidth available to their mobile devices. Compressed storage of signatures is vitally important.

There are several ways to compress signatures such that their size has a reasonable fixed upper bound and the performance of the localizer algorithm is unaffected by the compression. Our experiments have found that it takes only 100 representative scans, or approximately 2000 measurements (for the average of about 20 access points per scan we observed in our office building environment) to characterize the signature sufficiently for the localizer. Larger signatures add little if anything to localizer performance.

Another signature compression method which retains all available data is recording a histogram of signal strengths for each access point instead of a list of individual

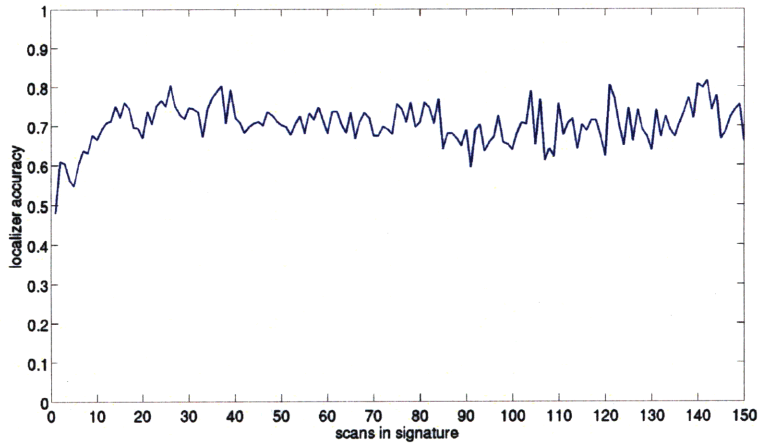

Figure 3-1: Localizer performance versus the maximum size of available signatures. On average 30 scans, or one minute of data, is the minimum signature size required.

readings. With the wireless drivers we tested, the resolution of the RSS values returned by the driver was approximately 1/100 of the range of the values. Thus, a histogram with 100 bins can capture the full shape of the signature for each AP at each location. Again, only 2000 numbers are required to represent the signature of a typical location. All of the algorithms described in section 6.1 can make use of the histogram compressed form of the signature, and some require even fewer numbers to characterize the signature, at the expense of accuracy.

3.3 Fetching signatures

Even when signatures have been compressed to roughly 2000 numbers per location, examining all of the available signatures in order to discover one's location would take an unacceptable amount of time. The building in which our experiments took place contains 1458 distinct locations. Even a small city contains millions of rooms and room-size areas for which signatures could be collected. One approach is to have the client transmit scans to a central server containing the available signatures, and have the server discover the client's location [15]. This takes advantage of the greater computing power available to a fixed server, but it exposes potential privacy concerns

because the user's location is being computed on a machine over which they have no control and transmitted over an open network.

For privacy reasons, we want to ensure that only the user's device knows its exact location. In order to do this, the mobile device needs a cache of signatures for places the user might be. Having a local cache also allows our location discovery service to operate during periods of lost connectivity. A mobile user only needs signatures for a few dozen locations in their immediate vicinity in order to cover their possible movements until their device can next connect to the signature server. How then to identify the best signatures to send to the client? What is needed on the server is a way to crudely determine the client's location without resorting to full location discovery.

One approach is to "cluster" scans [12] or signatures [50] offline based on the number of access points they have in common. This "AP commonality" is correlated with physical proximity, as shown in section 5.3. By sending one cluster of signatures to the client, one can ensure the client has signatures for an approximate neighborhood of spaces. If offline clustering of signatures is done well, the amount of data that must be sent to the client can be greatly reduced with be no loss of localizer performance compared to full global localization [37].

We find that explicit clustering of signatures is not required. Instead, we send the signature server a request to fill the client's local signature cache. This request contains a scan recorded by the client. Since a scan contains a list of MAC addresses currently visible to the client, the server can compute AP commonality between each signature in its database and the client's current location. If A is the set of MAC addresses in the scan sent by the client and B is the set of MAC addresses in a signature on the server, then the signature is likely to be near the place the scan was taken if:

$$\frac{|A \cap B|}{|A \cup B|} > \frac{1}{3} \tag{3.1}$$

This formula can be further refined by considering in the intersection of A and B only those access points where the signature contains some readings within $\pm\epsilon$ of the

reading in the scan. This criterion allows the server to select an initial set of signatures to transmit to the client which are likely to cover the client's physical neighborhood. Afterward, the client will keep track of how long signatures have been in its cache and will request updates when new data is likely to be available. The client will also periodically transmit a new scan and a list of signatures it currently has in its cache. If any locations not on that list show a sufficient AP commonality with the new scan, they will then be transmitted to the client as well. This minimizes the amount of data that must be sent to keep the client's cache up to date and covering the client's current neighborhood. This neighborhood is updated as the client moves, predictively incorporating locations the client might be in the near future. The extra time required to fetch these signatures is insignificant, since the computation is performed by a series of database operations on the server. These requests can be made anonymously to preserve user privacy [39].

3.4 Localizer inputs and outputs

The localizer algorithm requires a local cache of signatures and one or more recent scans to compare against them. For each cached signature, it calculates a likelihood that the recent scans match that signature. The most typical output to other routines would the signature with the most likely match.

Chapter 4

Experimental Data

We launched a test deployment of organic indoor location discovery in the Stata Center at MIT. About twenty Stata occupants were identified as candidate "test users" and each was provided a Nokia N810 tablet computer with prototype location discovery software. The test users were given short demonstrations of the software and instructed to "make a bind whenever you have been in a single place for a few minutes or intend to stay in a single place for a few minutes" [39]. The test deployment was confined to the Stata Center, an office building with ten floors and 1458 named spaces. Over the next twenty days, 16 users contributed signatures for 169 of those spaces. These signatures consisted of 640 distinct binds totalling over 117,000 scans and 17 million readings. These few users covered more than ten percent of a large building with less than a minute of effort each per day. We provided a mechanism to contribute binds for time intervals up to an hour into the past and future, and over half of all binds were over intervals. The remainder were "instantaneous" binds consisting of a few scans. The degree of contribution varied widely among users, as can be seen in the figure below. A few users (e.g. 9 and 13) contributed the majority of the bind-minutes. Because the N810 scans at a rate of 0.5 Hz, each bind-minute is equivalent to approximately 30 scans.

A prototype location discovery algorithm was part of the software provided. When the organic database was initialized, its accuracy was essentially zero since it had no available signatures from which to derive a location estimate. Over time its accuracy

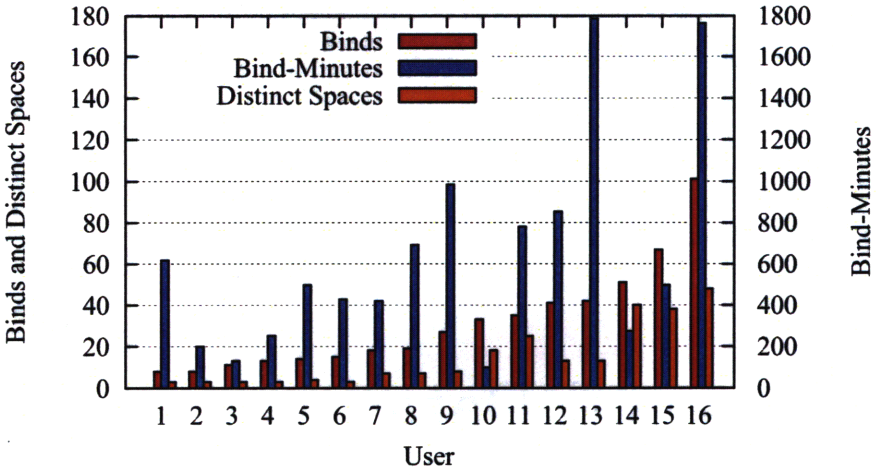

Figure 4-1: Signatures contributed by each user in an organic data collection experiment. Notable are users who contributed data for many spaces (15, 16); who contributed dense data for a few spaces (1, 9); and who contributed little data but benefited from others' contributions (2, 3).

improved, reaching 80% by the end of the experiment. The performance of the prototype localizer for our test users could only be determined when they specified bind intervals. During a bind the user is providing a location reference which we assume to be ground truth. We assessed performance by comparing the location the user specified with the localizer's location estimate for each scan during the bind interval. While the shortest "instantaneous" binds may have come from mobile users, longer interval binds generally came from users who were staying in one place for several minutes or more. Therefore the figures we obtained are most indicative of system performance for stationary users.

We further evaluated the accuracy of the prototype localizer by selecting 25 of the 169 spaces with available signatures and collecting new scans in each for 15 minutes. The localizer then attempted to discover the location of these new scans based on the existing corpus of organically collected signatures. For over 50% of these scans, the prototype localizer determined the correct location. For a further subset of them, it reported a location which was in the same room. We divided large rooms and hallways in our database into sections the size of a typical office. Distinguishing between such

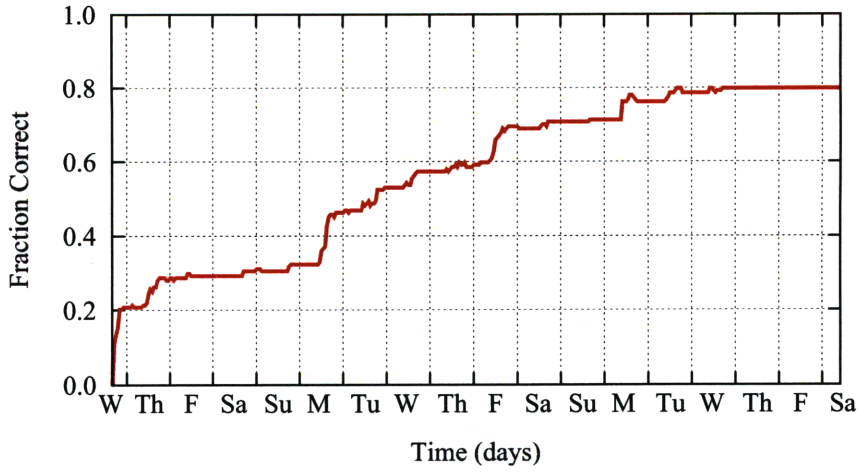

Figure 4-2: Performance of the prototype localizer algorithm over the course of the organic data collection experiment.

"room partitions" can be difficult because there is little to attenuate Wi-Fi signals between them, often just open air. Only 8% of the incorrect localizer estimates were places more than 10 meters from the true location.

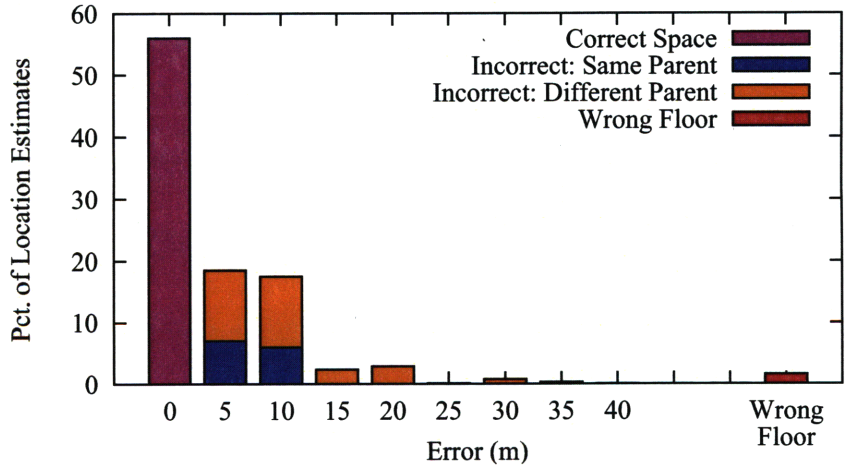

Figure 4-3: Prototype localizer algorithm accuracy in manual tests, broken down by correct/incorrect room and correct/incorrect floor.

Chapter 5

Signal Characteristics

The signals used by Wi-Fi networks are transmitted on an open frequency band at 2.4 GHz. Other devices, such as some cordless phones, also transmit in this band and cause interference in Wi-Fi signals. Microwaves can generate noise at this frequency, and 2.4 GHz radiation is absorbed by water molecules. Humans, being in part large concentrations of water, cast radio frequency shadows in Wi-Fi signals. Walls and furniture both attenuate and reflect the signals, causing them to follow multiple paths from transmitter to receiver [23].

The factors affecting Wi-Fi signals can be broken into three groups. Walls, doors, furniture, and other fixed features create a complex but essentially static baseline pattern of received signal strength which can vary measurably between places only tens of centimeters apart [49]. Cordless telephones, microwaves, and other occasional sources of interference could cause temporary degradation of an indoor location service that uses Wi-Fi signals, but are generally sporadic enough to not need consideration. People cause a complex and moving pattern of RF shadows which will affect transmissions from some directions but not others. The shadow cast by the user on their device attenuates signals by approximately 5 dB. This is strong enough that given enough training samples, one can determine which direction the user is facing from the RSS values their device records [3].

5.1 Time of day variation

The attenuation caused by people walking through the radio environment tends to follow set daily and weekly patterns. The charts below show received signal strength histories for 24-hour periods during the weekday and weekend in a busy office corridor. The afternoon period when occupants of the office are most active are clearly visible, especially for the access point with the strongest signal. Anyone entering or leaving the office would walk through the direct signal path from this access point to the receiver, casting an RF shadow. [The RSS values in figures 5-1 through 5-4 are on a different scale than in other figures in this thesis. That is because these 24-hour observations were recorded with a different wireless card. Other figures use the RSS scale reported by the Nokia N810 tablet unless otherwise noted. See section 7.3 for a discussion of calibrating RSS scales between different wireless cards.]

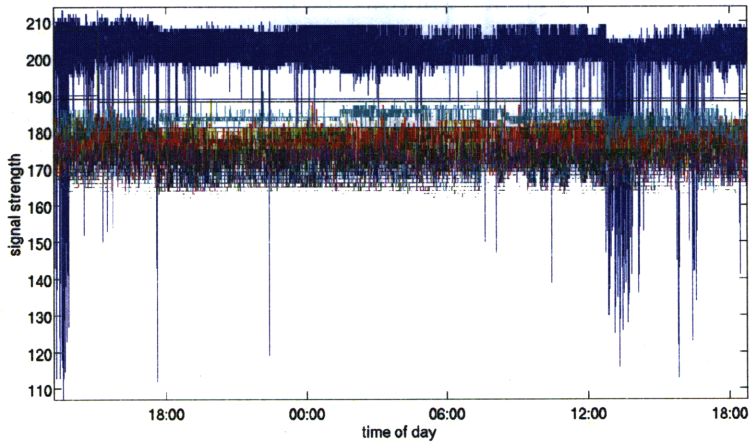

Figure 5-1: Variation in RSS for a fixed receiver over 24 hours on a weekday.

Offices, homes, retail businesses, and other types of locations will show different, but largely predictable, time-of-day variations in the local wireless signature due to the presence of people. If all the available training data was collected at a time when the area was nearly empty, the localizer algorithm may exhibit impaired performance during busy hours [22].

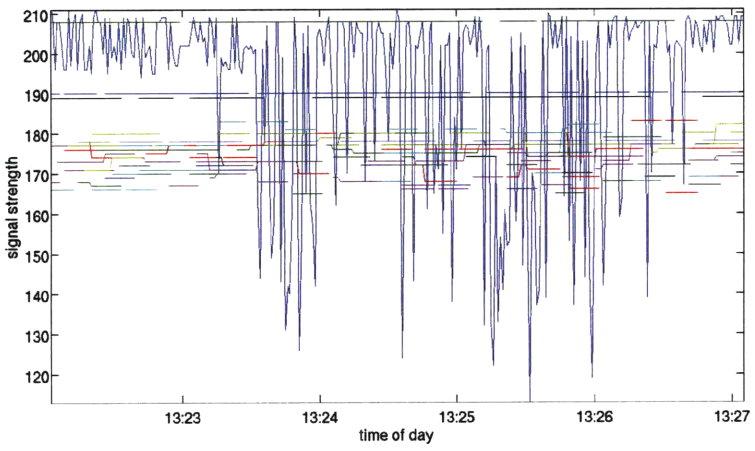

Figure 5-2: Detail: five minutes of scans from weekday data.

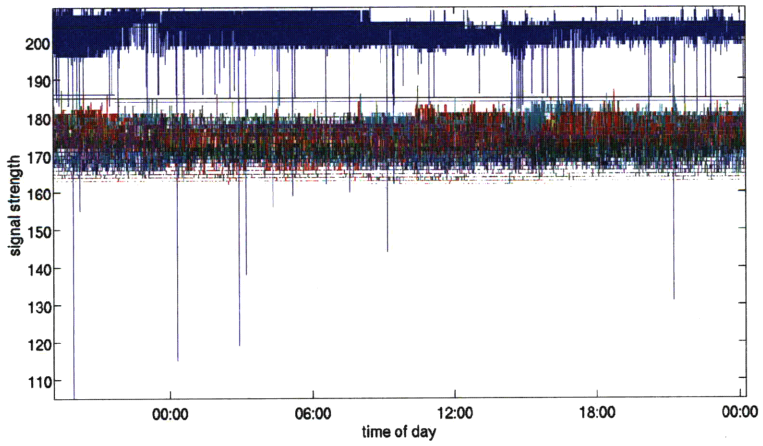

Figure 5-3: Variation in RSS for a fixed receiver over 24 hours on a weekend.

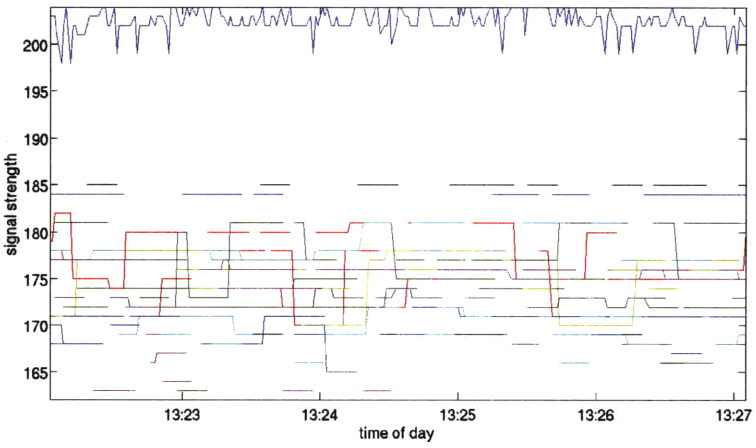

Figure 5-4: Detail: five minutes of scans from weekend data.

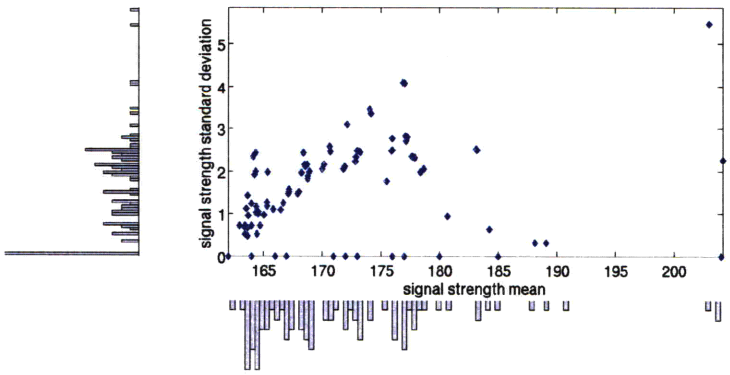

Figure 5-5: Relationship between mean and standard deviation of received signal strength by access point. In general, stronger signals display more variance in RSS. Data set is the same as for figure 5-1.

5.2 Network changes

When a location discovery service is deployed over a long time period, changes occur even in "fixed" parts of the configuration. Walls and furniture may be moved or changed to different materials. Some access points may be replaced by others with a different MAC address, or the physical arrangement of access points could be changed. Such changes are usually local and happen in a piecemeal fashion. Survey-based location methods would require a re-survey in such circumstances. Organic collection methods are continually re-surveying. Most changes to the physical environment or the network would manifest themselves as a region of degraded localizer performance, which can cue an organic system to request new data from users in the area. Attempts have also been made to adapt to network changes using a neural network model [1].

5.3 Useful properties of signal patterns

There are several statistical properties of received signal strength measurements from a Wi-Fi network which are of interest to designers of a location discovery system. Despite the effects of intervening walls and multipath fading, the received intensity is still correlated to distance from the transmitting access point. Higher values are recorded near the access point and lower values further away, as expected. Wireless drivers generally have a threshold below which signals cannot reliably be received. At any location, access points which have an RSS near this threshold appear in only a fraction of scans taken at that location. Nearer access points or those with a higher RSS are seen more reliably.

Finally, because each access point has a finite neighborhood in which it can be seen by a wireless card, the number of access points that two locations have in common is correlated with the distance between them. This "AP commonality" will be important when designing a localizer algorithm which is robust to the varying signature sizes returned by organic data collection. Even in the absence of RSS information, the AP commonality between two scan records can contain enough information to

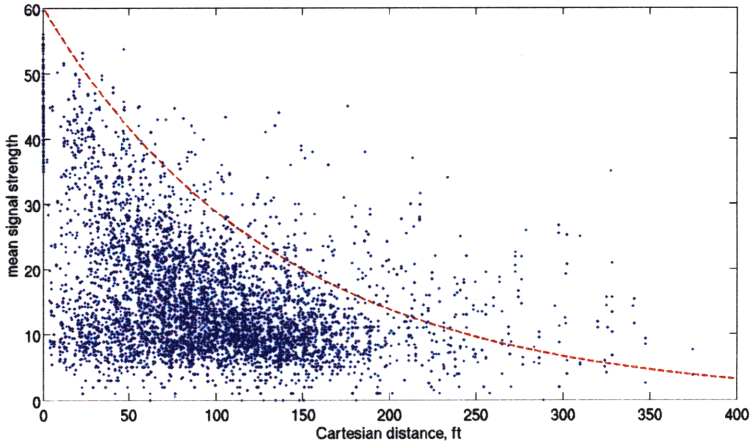

Figure 5-6: Relationship between mean RSS and distance to the transmitting access point. The data broadly agrees with the exponential decline with distance predicted by simple signal propagation models.

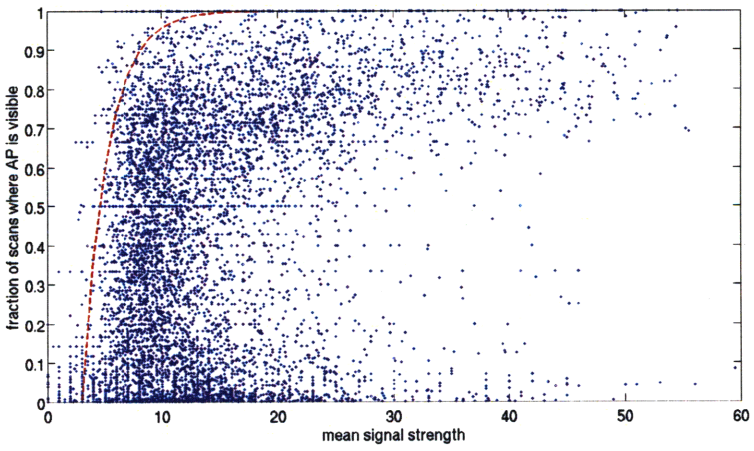

Figure 5-7: Relationship between mean RSS and the fraction of scans in which an access point is seen. A roll-off at low RSS values is clearly visible.

determine whether two devices are in the same room [5]. Therefore this is a powerful supplementary metric which one should bear in mind when designing a localizer algorithm.

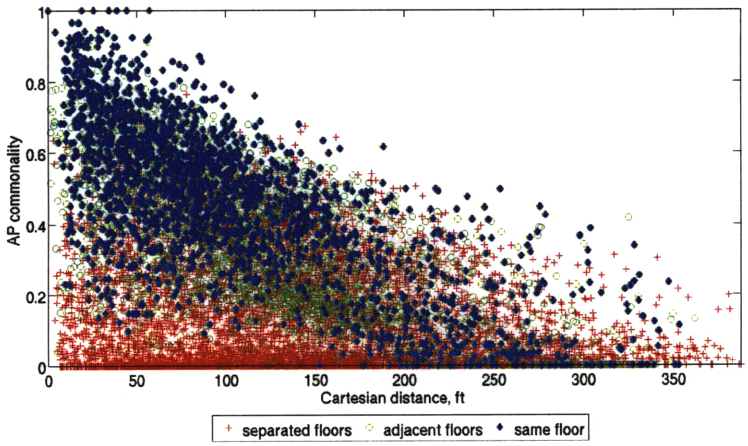

Figure 5-8: Relationship of distance between rooms and the access point commonality between their signatures. The relationship is strongest for pairs of locations on the same floor.

Section 3.3 described a method for finding signatures from locations near the location of a reference scan by setting a minimum threshold of AP commonality. Figure 5-9 shows the effect of such a filtering criterion on pairs of signatures. As the threshold is tightened, more of the remaining signature pairs are from locations which are physically near each other. This demonstrates that in practice the AP commonality threshold achieves its stated purpose in section 3.3 by eliminating signatures of locations which are far from the reference.

5.4 Correlation of access points

There are a few other important properties of Wi-Fi signal transmission which should be examined in order to build an effective localizer. The first is the correlation between signals transmitted by different access points. If one were to plot the signal paths from an access point to a fixed receiver in an indoor environment, the sequence of reflections and attenuating objects the signal must pass through would depend heavily on the

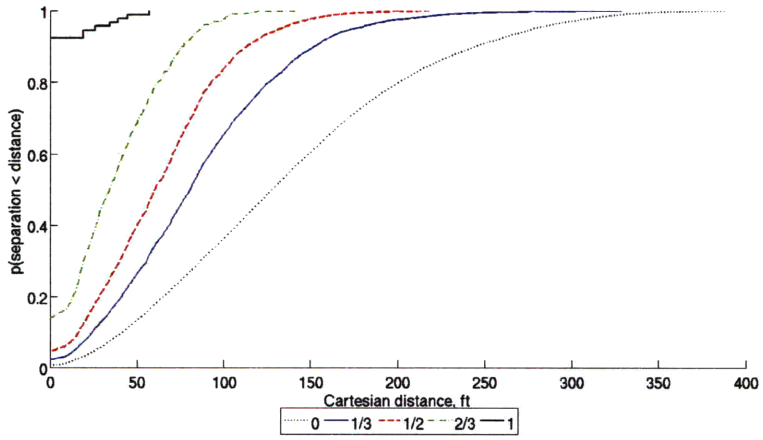

Figure 5-9: Distribution of distances between rooms whose signatures meet different minimum AP commonality thresholds.

location of the access point and the location of the receiver. A deployed wireless network rarely has access points physically near each other; they are distributed around the building in order to maximize network coverage. As a result, from the point of view of a receiver inside the building signals from nearby access points should be arriving from different directions. This means that the transitory effects shown in section 5.1 are unlikely to affect multiple signals in the same way at the same time. The result is that for a fixed receiver, the signals from different access points are uncorrelated, as shown in the figure below. The localizer designs presented in section 6 make use of this convenient property.

Correlation in the time domain is also an important design consideration. Figure 5-11 shows the average autocorrelation observed by a fixed receiver. Over short time periods, this correlation means that scans cannot be treated as independent samples. Most importantly, using a simple average of the RSS values from several consecutive scans is not as accurate as averaging the location estimates produced by each scan separately [46]. This is a result of a relatively stable radio environment, the property which makes Wi-Fi-based localization possible. There are several ways to deal with correlation between scans. One can attempt to model it explicitly [49], or consider scans individually and combine the information after localizing each one [26].

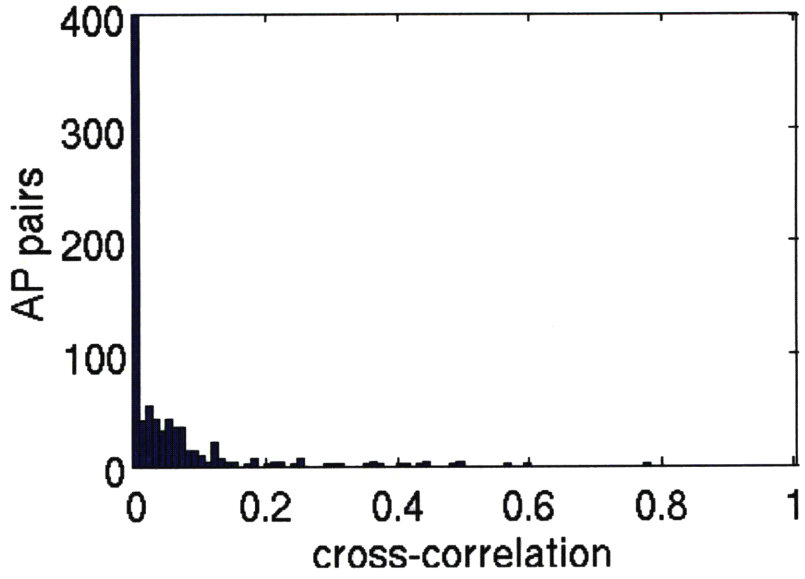

Figure 5-10: Correlation between access points' RSS for a fixed receiver. No significant correlation is evident, suggesting that access points' RSS values are independent given the receiver's location. Data set is the same as for figure 5-1.

The latter approach will be discussed in section 6.3.

5.5 Signal dispersion

The distribution of RSS values observed for a single access point at a single location may take on a variety of shapes (see figure 6-1 for one example). However, if one looks at the dispersion of the RSS, specifically the distribution of differences between RSS observations, a pattern emerges. When averaged over multiple access points and multiple locations, the signal dispersion takes on the shape shown in figure 5-12 regardless of the specific receiver hardware or software. An exponential curve can be fitted to this data. This provides an estimate for the expected difference between readings $E_{i \neq j} [|o_i - o_j|]$ for a typical access point seen from a typical location. This estimate proves useful for filling in missing data in wireless signatures in section 6.1 and as a way to calibrate unknown wireless devices in section 7.3.

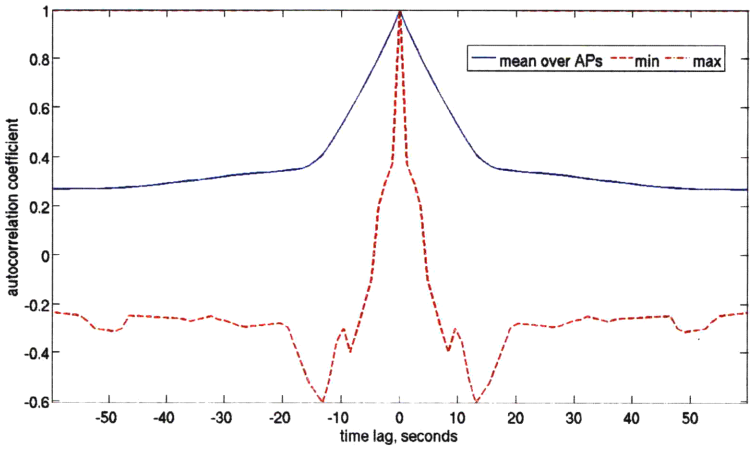

Figure 5-11: Autocorrelation of RSS for a fixed receiver, averaged across access points. Correlation between readings less than 15 seconds apart is significant. Data set is the same as for figure 5-1.

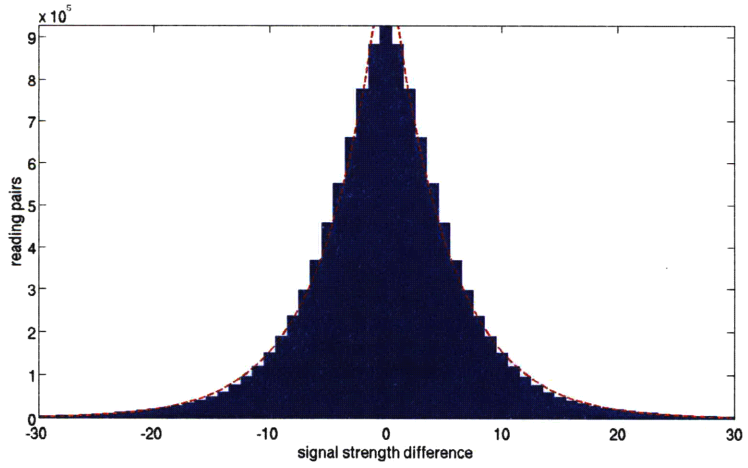

Figure 5-12: Dispersion distribution for RSS measurements, with fitted exponential approximation.

Chapter 6

Algorithm

The task of a Wi-Fi based location discovery system is to determine the location of a wireless-enabled device given a scan or set of scans collected by that device and the signatures which have been collected to date. The user is assumed to be in one of the locations represented by the signatures, and the job of the localizer algorithm is to determine how likely it is that each location is the one which contains the user. This is a natural application for a Bayesian classifier. If the location of the user is l, a variable whose domain is the set of named locations for which signatures are available; and the scan returned by the wireless card is represented by observations o, then:

$$p(l|o) = p(l)\frac{p(o|l)}{p(o)} \tag{6.1}$$

Several assumptions can be applied to simplify this formula. Access points are normally found distributed throughout a building, so the signal from each must travel along a different path to reach the sensor. Therefore it is reasonable to assume that measurements from different APs are uncorrelated, given the location of the sensor. Figure 5-10 indicates that these measurements are observed to be uncorrelated in real data, so the assumption is justified. With this assumption of conditional independence, the formula becomes that of a naive Bayesian classifier.

$$p(l|o) = p(l)\prod_{AP_i}\frac{p(o_i|l)}{p(o_i)} \tag{6.2}$$

Naive Bayesian classifiers are easy to construct, so this is a very convenient set of assumptions. Unfortunately, this formula only works well only when each signature contains an approximately equal number of scans and each scan contains an approximately equal number of readings. Both of these conditions are frequently and widely violated in actual practice, especially when signatures are collected organically. Correcting for these severely unequal inputs is the central design challenge for an organic indoor localizer algorithm. Section 6.1 characterizes the likelihood $p(o_i|l)$ given the recorded signature for access point i at location l. Section 6.2 deals with the deceptively challenging process of combining evidence from different access points when the number of access points per signature varies. Section 6.3 discusses different approaches for selecting the location prior probabilities $p(l)$.

6.1 AP - level comparison

If a signature contains readings from a MAC address which also appears in the scan being localized, then the task of the lowest level of the algorithm is to determine the probability that the signature's readings were taken in the same local radio environment as the scan's reading. The probability $p(o_i|l)$ is the likelihood that the data in the scan and the signature for AP i were drawn from the same distribution.

6.1.1 Histogram

One way to characterize the distribution is to use the signature's data directly. The recorded readings in the signature are binned into a histogram and the likelihood of a match is the count for the bin in which the scan's reading o_i falls divided by the signature's total count of readings. This technique was used with some success in early versions of the Rice Wireless Locator [24].

$$p(o_i|l) = \frac{n(o_i)}{n_{total}} \tag{6.3}$$

The obvious limitation to using the histogram directly like this is that a signature may not contain enough scans to accurately characterize the distribution. The histogram could contain unrepresentative gaps and unevenness. As a consequence, the results of the probability computation are not likely to be accurate.

6.1.2 Normal distribution

Another approach is to fit a distribution to the signature data. The Rice group improved their accuracy and reduced their signature storage requirements by fitting signatures to the normal distribution [26]. In this case, only the mean and variance of the fitted distribution need to be transmitted to the user and the probability is computed from a normal probability density function, N.

$$p(o_i|l) = N(o_i, \mu_i, \sigma_i) \tag{6.4}$$

Unfortunately, the observed distributions often take on a distinctly non-normal shape due to multipath effects and interference.

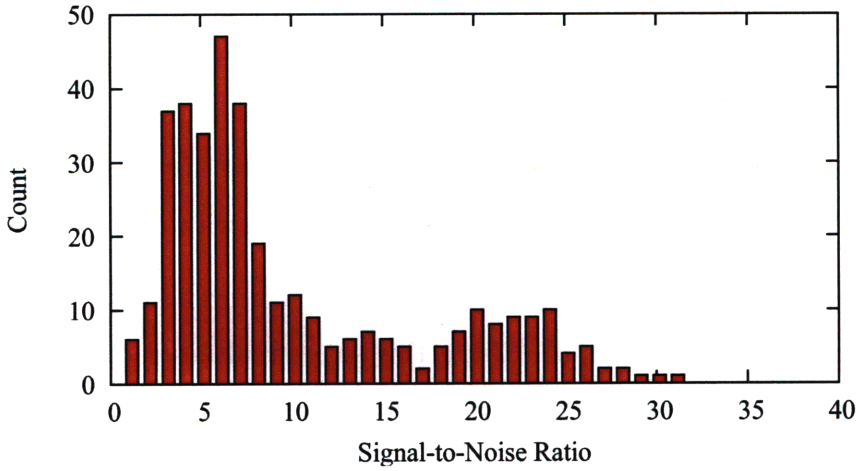

Figure 6-1: Example RSS signature for one access point showing non-normal shape.

6.1.3 Student t-test

A refinement of the normal-fit approach which we investigated was fitting a Student's t distribution to the data [41]. This resulted in a small performance improvement over a normal fit. The main advantage of this approach is that optimized routines exist for computing $p(o_i|l)$, the probability that both samples were drawn from the same distribution, by Student's t test. However, this method also assumes that the underlying distribution is normal.

6.1.4 Smoothed histogram

The distribution represented by the signature could be fit much better by a mixture of gaussians or other distributions. This introduces the complication of determining how many modes are represented in the data, however. All such fitting is likely to distort the shape of the distribution somewhat. In our experience, there isn't a pressing need to compress the signatures to that degree to have a functional system capable of fetching signatures from a central server as needed. Given that the ratio between the dynamic range of values reported by most wireless drivers tested divided by their resolution is approximately 100, the histogram form is a suitably compact representation of a signature. The uneven sampling which is the histogram method's main weakness can be addressed by convolving the histogram with a smoothing filter [24].

A good candidate for the filter kernel is the exponential shape seen in the dispersion of readings in section 5.5. Using this shape replaces each reading with the average expected distribution of readings, given that one. In effect, this uses the properties of the sensor to approximately fill in missing samples. Each reading o_{sig_i} in the signature contributes to the match probability if it is sufficiently close to o_i. In our design, rather than convolve the histogram with this smoothing kernel before computation, we retain the original histogram and perform the convolution on-line,

in the following way:

$$p(o_i|l) = \frac{1}{n_i} \sum_{o_{sig_i}} \begin{cases} 1, & |o_i - o_{sig_i}| \leq \text{mindiff} \\ exp(k|o_i - o_{sig_i}|), & \text{mindiff} < |o_i - o_{sig_i}| \leq \text{maxdiff} \\ 0, & |o_i - o_{sig_i}| > \text{maxdiff} \end{cases} \tag{6.5}$$

We used 2 for the minimum difference mindiff and 25 for the maximum difference maxdiff. By smoothing the observed distribution while retaining its general shape, this method out-performs both the simple fitting approaches and the direct use of the histogram. Its computational load is higher than approaches which perform compression offline but it allows different clients to use different kernels tailored to their wireless driver. As figure 6-2 shows, the smoothed histogram provides a much better fit than a normal distribution does to RSS distributions with arbitrary shapes.

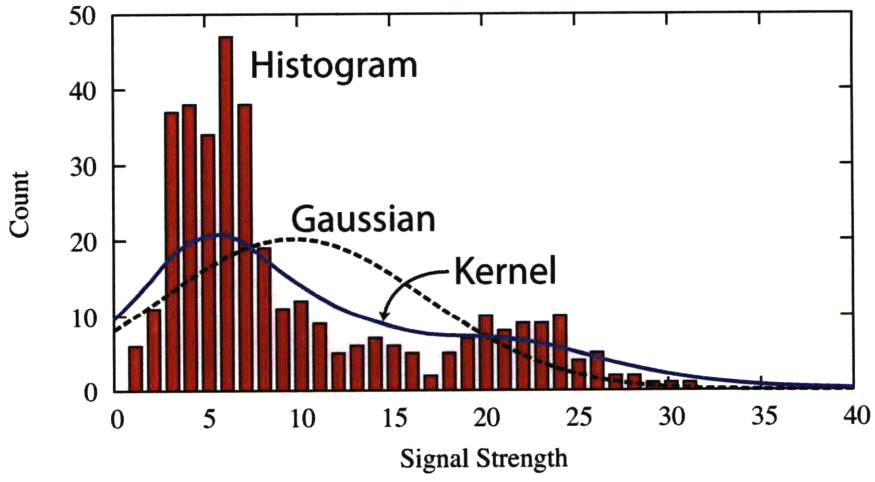

Figure 6-2: Normal and smoothed histogram fit to an example RSS signature for one access point.

6.1.5 Presence weighting

An enhancement on any of these approaches is to weight the match probability from each access point by how frequently that access point appears in the scans which comprise the signature. This reduces the effect of APs which appear sporadically because they are near the wireless card's sensing threshold. An AP can be heard sporadically over a much larger area than it can be heard consistently, so the consistent APs contribute more reliable information. If access point i is visible in n_i scans out of a total N_l scans for location l, then the weighting adjustment for that access point would be:

$$p'(o_i|l) = \frac{n_i}{N_l} p(o_i|l) \tag{6.6}$$

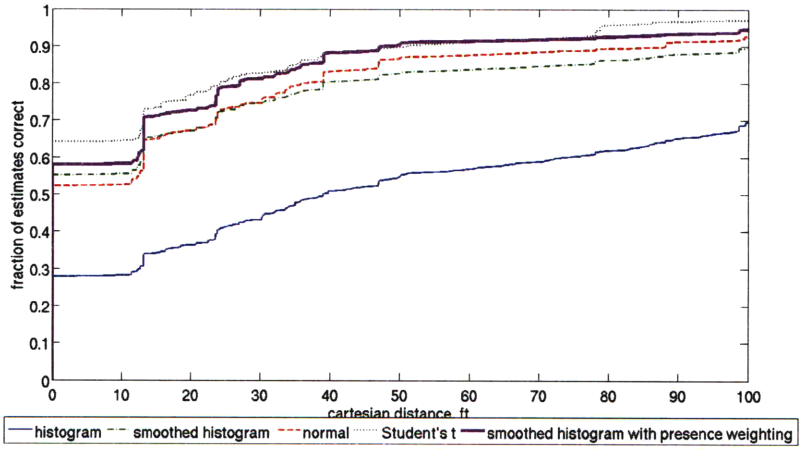

Figure 6-3: Performance comparison of different AP-level matching methods.

6.2 Room - level comparison

6.2.1 Combination by "AND"

When localizing a scan with readings from M access points, each location considered will have $m_l \leq M$ access points in common with the scan. One of the formulae in the last section can be used to obtain a match probability for each of these APs. A naive Bayesian classifier design as described in section 6.2 suggests that the way to

create an overall probability for each location is to multiply together the AP match probabilities for that location.

$$p(o|l) = \prod_{AP_i} p(o_i|l) = p(\text{AND}_{AP_i}(o_i|l)) \tag{6.7}$$

Unfortunately, this approach works well only if all locations have approximately the same AP commonality with the scan. Other localizer designs have attempted to enforce the criterion that m_l be approximately constant for all locations either by eliminating access points from consideration [26] or examining only those locations with the highest values of m_l [52]. The reason why such precautions need to be taken is obvious. Consider two locations A and B. The signature of A has three access points in common with the scan, and each returns a match probability of 0.7. The signature of B has two access points in common with the scan, and each returns a match probability of 0.6. The overall $p(o|l)$ for B would be 0.36, yet for A it would be only 0.343! Both the greater number of common APs and the greater match probabilities for A are correlated with shorter error distances, as we saw in section 5.3, so clearly the Bayesian approach returns overall probabilities in the wrong order in this case. On real data, using AND to combine AP probabilities without controlling for differences in m_l results in nearly 0 accuracy, worse than random! Therefore we will consider different approaches for combining AP match probabilities which are more robust to varying signature sizes and which rank locations correctly based on the properties of signatures which we know are related to location.

6.2.2 Geometric mean

The failure of combining AP probabilities by AND (without compensatory mechanisms to level the playing field) happens because the larger m_l is, the more probabilities between 0 and 1 are multiplied together, and the smaller the result becomes. An obvious way to counteract this tendency is to use the geometric mean of the AP

probabilities instead of the product.

$$p(o|l) = \prod_{AP_i} p(o_i|l)^{1/m_l} \tag{6.8}$$

This normalizes the result by m_l and gives an improvement in performance.

6.2.3 Combination by "OR"

The more AP commonality there is between the signature and the scans being localized, the closer the wireless device is likely to be to the signature's associated location, as noted in section 5.3. Therefore, it would be nice to have a method of combining AP probabilities that actually increased as m_l increased to favor locations with higher commonality. One candidate is to use probability that any reading matches the signature, instead of the probability that all readings match the signature. By including the likelihood of matching subsets of the observed APs, this probability grows closer to 1 as m_l increases.

$$p(o|l) = 1 - \prod_{AP_i} (1 - \gamma p(o_i|l)) = p(\mathrm{OR}_{AP_i}(o_i|l)) \text{ if } \gamma = 1 \tag{6.9}$$

The parameter $\gamma \in [0, 1]$ is used to keep the resulting probabilities from getting too close to 1, which can confuse the ranking if differences between locations get too close to the machine precision. Also, smaller values of γ reduce the differences between AP match values by driving them closer together, placing more weight on AP commonality and less on the individual AP matches. A good compromise value for γ tends to be between 0.1 and 0.3.

6.2.4 AP voting

There's another information inequality to consider at this stage of the algorithm. Some readings in the scan may have relatively high values. High readings only occur in a few locations near the access point. Lower readings, by contrast, can be seen in a large perimeter of spaces further from the access point. Therefore, if one looks at the

distribution of $p(o_i|l)$ over the locations considered, higher o_i tend to have peakier distributions focused in just a few rooms and lower o_i tend to have distributions spread over many rooms. The higher reading will match well in a few places, providing useful information. The lower reading will match well in many places, providing less useful information. As yet we have not distinguished between these situations. Other researchers have found that matching only the 3 or so APs with the highest recorded RSS in the scan produces similar results to considering all APs [52].

What if each access point in the scan had one vote to distribute across all the locations? Peakier distributions would concentrate their votes in a few locations and more spread-out distributions would dilute their votes across many locations.

$$p'(o_i|l) = \frac{p(o_i|l)}{\sum_l p(o_i|l)} \tag{6.10}$$

$$p(o|l) = \sum_{AP_i} p'(o_i|l) \tag{6.11}$$

This simplest form of an AP voting algorithm performs as well as or better than the other combination methods which have been discussed here. Its only significant drawback is that it requires retaining AP match probabilities for all locations in order to perform the normalization step. Possible refinements include giving high readings a larger total vote than low ones, or combining the normalized probabilities in a way other than a simple sum.

6.3 User motion models

The distribution of prior belief over locations is a component of a localizer algorithm which often receives little consideration, though the effects of some different prior distributions have been studied theoretically [51]. The simplest approach to determining the prior probability over locations $p(l)$ is to consider all locations equally likely. This is a common tactic when performing global localization given no prior knowledge [22]. If the number of locations in the available database is large, then for efficiency's sake it would be helpful to consider only those whose signatures have a relatively high AP

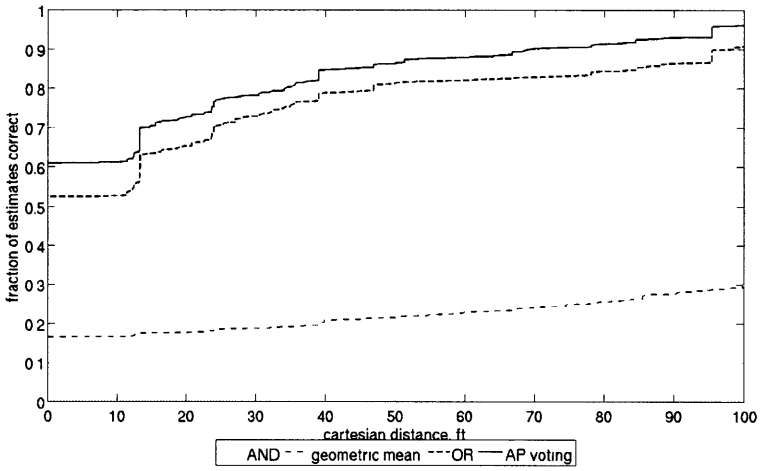

Figure 6-4: Performance comparison of different AP combination methods.

commonality with the scan being localized [50]. Our algorithm uses the local cache to perform this restriction implicitly (see section 3.3).

For tracking mobile users, additional information is available which can be incorporated into the prior. We make use of scripts which read dxf floor plans generated by AutoCAD and extract the locations of walls, doorways, and open-air connections between rooms (such as the transition between adjacent sections of a corridor). Using these, we can build a graph of paths a mobile user could take from one room to another without walking through walls. The Rice Wireless Locator group used a similar graph in a Hidden Markov Model to improve their algorithm's accuracy for moving users from about 50% to over 70% [26]. They assigned a prior probability of 0.7 to the previous position solution and divided the remaining 0.3 evenly among its neighbors. A more sophisticated approach would be to use the full posterior probability calculated for scan t-1, combined with similar transition probabilities, as the prior probability for scan t and use the Viterbi algorithm to track multiple hypotheses of user motion. This remains to be tested.

Whether the model starts from the most likely previous location or tracks all likely previous locations, the transition probabilities between locations are simply described: some positive δ if two rooms are connected, and 0 if they are not. This approach does not take into account the relative popularity of different exits from a

room. The main door to the hallway may be used often, but the door to the closet may not. Furthermore, if no graph of the interconnections between locations exists, no HMM approach can be used. It would be desirable to increase or decrease the weights on different edges of the location graph based on how often users traversed those edges. Collecting such information organically may be impractical, since it would require users to specify not only where they were but what paths they took. Using a history of localizer estimates to identify which path a user took is feasible, but involves potential complications because the results of the algorithm will be fed back into its decision-making process. If one can determine the relative popularity of the connections between a location and its neighbors, then the HMM can assign transition probability proportional to that popularity. The weights are also useful to route-finding algorithms, which could use them to direct users along popular (and presumably easier to follow) paths.

Other approaches for modeling user motion include using a particle filter populated with hypothetical user paths [31]. In grid-survey applications, there has also been success with using a weighted average of the 3 or so locations with the highest probability, or an average of the most recent top matches [47]. This averaging approach is arguably more effective than a nearest-neighbor match for grid surveys [48]. A user's device could also privately record their typical movements and make predictions without a record of which locations are physically connected [4]. This would be useful where full floor plans are not available.

Chapter 7

Future Directions

7.1 Incorporation into Rich Maps

The software tools which allow organic collection of wireless signatures can be extended to allow users to contribute any sort of data. Our group is currently developing this sort of application, which we call a "rich map." Rich maps will enable map annotation and a geotagging-like capability for data indoors, with room-level accuracy, and with a selection of different sharing models. One advantage of incorporating rich maps into organic data collection is that instantaneous signatures can be collected each time a user tags an annotation or other data to a location. Organic collection can then proceed silently in the background as the user interacts with a more compelling application. By encoding information about room uses and the connections between rooms, a rich map can enable route-finding indoors and support location-based search. Indoor location discovery is a vital background component of both of these services.

7.2 Validation of contributors' data

A major challenge of using organically collected data for location discovery is that not all users accurately identify their locations when they contribute scans to the signature database. Some will be simply mistaken, others may be maliciously attempting to

mislead the system. To create a trusted, widely-distributed location discovery service one needs a way to vet newly contributed data. If a user contributes a bind to be added to an existing signature, then it would be expected to match that signature with a high likelihood. Running any effective localizer algorithm on new contributions would identify those which may not be correct. Sometimes poorly matching new scans are correct, however. If there has been a change to the network the old signature may not apply anymore. That is why it would be useful to retain binds which match poorly but flag them as a "quarantined signature." Over time, more binds will become available. If they match the quarantined signature, a change in the network is a more likely hypothesis. The quarantined signature would be reinforced and eventually it will supplant the previous signature. If additional binds reinforce the previous signature, the quarantined signature can eventually be removed by a garbage-collection routine.

Certain types of errors will still be difficult to detect. Distinguishing between adjacent rooms can be difficult for a localizer [26]. If users mistakenly attribute binds to a room adjacent to their true location, the error may go undetected. The first users to contribute signatures for an area have to be trusted until their data is either reinforced or overruled by subsequent users. Finally, a sufficiently persistent and dedicated malicious contribution of false data could still prevail, but that is a problem common to all services relying on user-contributed content.

If user contributions are being filtered, one can consider filling gaps in the organic database with interpolated signatures. This has been tried with some survey-based indoor localizers [49]. When survey points follow a dense grid, missing data can be interpolated from adjacent points, but it is not clear whether this approach would work for room-level signatures [30]. Merely extrapolating missing signatures from data collected nearby could at least allow a user's localizer to consider the unvisited locations. This extrapolated data would be less reliable than actual readings taken by users, so users should be encouraged to replace it with real data and it should be immediately replaced when binds become available.

7.3 Calibration of new devices

All wireless device drivers report some measure of received signal strength from visible access point, but it seems no two of them use the same scale when reporting it. The group which designed the Rice Wireless Locator discovered that the values reported by the drivers they tested could be related with a simple linear correction [22]. This holds for the drivers we tested, as well.

$$RSS = a + b * RSS_{reference} \tag{7.1}$$

For each driver, one needs to find an offset and a scaling factor for its reported scale. The approach used by Rice is to take an uncalibrated device to a location with a surveyed signature and collect a signature using the new device. Given these two sets of readings, they would then determine the maximum likelihood values for the two calibration parameters [26]. This is a good approach, but it does require access to a location which has a recorded signature and which the user of the new device can identify. One of the properties identified in section 5 can be used to simplify the process. The dispersion of readings as seen in figure 5-12 can be fit to an exponential curve. The scale factor b will affect the width of the exponential peak and can be determined from a sufficient number of scans regardless of where they were collected, as long as the user stays relatively stationary. The offset a cannot be determined from this information, but a localizer could be designed to simultaneously solve for location and the offset factor.

7.4 Incorporation of accelerometer data

Some mobile devices such as the Nokia N95 and other smart phones and PDAs incorporate low-quality accelerometers. Though not suitable for dead-reckoning navigation, these accelerometers have been used for applications that identify the user's level of physical activity and even count steps with some accuracy [21]. Identifying intervals of time when the user is stationary can be useful when organically collecting

data. Scans can be collected and stored until the user can be asked to volunteer where they were during each period of time. If accelerometer data can provide a coarse estimate of how quickly the user is moving, that information can be used to refine the localizer's estimate of the user's instantaneous location, as described in section 6.3. Accelerometer data can also filter possible user locations by matching the sensed activity (sitting, standing, walking, etc.) against the activities of previous users in each location [33].

7.5 Incorporation of GPS and other sources

Many mobile devices which can connect to Wi-Fi networks also have embedded GPS receivers. Merging the results of GPS localization and Wi-Fi localization would seem to provide a seamless location discovery solution which performs better than either method alone. There are some obstacles to performing this combination. Indoors and outdoors near buildings, significant portions of the sky are blocked, so signals from GPS satellites are difficult to receive and unreliable due to multipath fading. The vast majority of Wi-Fi access points are located indoors, and their signals propagate outdoors some tens of meters [16]. This places the region of coverage overlap between Wi-Fi and GPS outdoors near (but not too near) buildings. Potentially, both GPS signals and Wi-Fi signatures are degraded in such locations. A greater challenge arises from GPS reporting its results in geodetic coordinates and a Wi-Fi localizer reporting its results in terms of defined locations. Combining the two requires assuming an error distribution for the GPS solution and determining how much of it lies within each location considered by the Wi-Fi localizer.

An alternative use for combined GPS and Wi-Fi localization would be to "anchor" indoor maps to georeferenced coordinates. Once such anchors are established, a position solution from GPS would provide a way to "bootstrap" the Wi-Fi localizer's signature cache by allowing it to identify nearby locations. A method for performing such "bootstrapping" with a wireless scan was described in section 3.3. Another method would be to examine the user's daily calendar or schedule and fetch signatures

for locations near the places the user is supposed to be in the near future.

7.6 Mapless location discovery

Despite taking advantage of user contributions to build wireless signatures, our system still requires an underlying map be provided. In order to expand service outside of areas with public, managed floor plans, users must be empowered to create their own maps. The WikiMapia project uses an online community of millions of users to draw boundaries around locations shown on Google Maps and label them with text, photos, or videos [45]. Signature-based location discovery methods can use user-generated maps even if they are not precise, as long as they are notionally accurate and correctly labeled.

Unfortunately, user-provided labels can be vague, with their level of specificity dependent on the intended audience. Intel's PlaceLab group found user annotations ranging from "living room" to "Canada" [13]. In context, these labels are still potentially useful, even when official room names are available. A user could build a database of idiosyncratically labeled "personal signatures" of places they commonly visit without any reference to a map, and it would still useful in predicting connectivity between these places and where the user is likely to go next [4].

Chapter 8

Conclusion

This thesis attempted to synthesize the extant literature on indoor location discovery from Wi-Fi networks and to directly compare a representative selection of localizer algorithms. We have attempted to note the useful features and challenges that Wi-Fi presents for location discovery. It is our intent to provide the reader with enough information to design an effective indoor localizer algorithm. However, some caution is required.

The wireless networking standards were not designed with location discovery in mind, so many features are poorly designed for the task. Wireless card drivers have no common standard for reporting RSS, though all seem to be related to the dB attenuation of the signal. Some drivers do not report all visible APs, or they update the RSS of some or all APs at a rate that is too slow for effective localization of a mobile user. Furthermore, the upcoming 802.11j standard allows access points to dynamically adjust their transmission power and channel, introducing potentially difficult real-time calibration challenges for Wi-Fi location discovery systems. Finally, it is worth bearing in mind that Wi-Fi technology has become widespread only in the past ten years and may be supplanted just as quickly.

The most enduring message to be taken from these investigations is the difficulty of discovering location using a radio map that is largely unknown, subject to occasional significant changes, and affected by significant dynamic noise. Key considerations include distributing the mapping process using organic signature collection, dividing

areas into regions which are both meaningful to users and clear enough to be shared, and designing localization algorithms which can perform well despite significant disparities in data density and incomplete sensing.

Bibliography

[1] Aylin Aksu, Joseph Kabara, and Michael Spring. Reduction of location estimation error using neural networks. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[2] Bharath Ananthasubramaniam and Upamanyu Madhow. Cooperative localization using angle of arrival measurements in non-line-of-sight environments. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[3] Paramvir Bahl and Venkata N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. *Proceedings of InfoCom 2000*, 2:775–784, 2000.

[4] Ingrid Burbey and Thomas Martin. Predicting future locations using prediction-by-partial-match. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[5] Alessandro Carlotto, Carlo Bonamico, Fabio Lavagetto, Massimo Valla, and Matteo Parodi. Proximity classification for mobile devices using Wi-Fi environment similarity. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[6] Yu-Chung Cheng, Yatin Chawathe, Anthony LaMarca, and John Krumm. Accuracy characterization for metropolitan-scale Wi-Fi localization. *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services (MobiSys 2005)*, pages 233–245, June 2005.

[7] Jakob Eriksson, Hari Balakrishnan, and Samuel Madden. Cabernet: Vehicular content delivery using wifi. *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking (MOBICOM'08)*, pages 199–210, September 2008.

[8] Nokia Europe. Nokia Maps 2.0 website. `http://europe.nokia.com/A4509271`, 2009.

[9] Google Maps. http://maps.google.com/, 2009.

[10] Locale group. Android Locale website. http://www.androidlocale.com/, 2008.

[11] Andreas Haeberlen. Rice Wireless Locator source site. http://www.cs.rice.edu/~ahae/rwl/, 2007.

[12] Jeffrey Hightower, Sunny Consolvo, Anthony LaMarca, Ian Smith, and Jeff Hughes. Learning and recognizing the places we go. *Proceedings of Ubicomp 2005*, pages 159–176, September 2005.

[13] Jeffrey Hightower, Anthony LaMarca, and Ian Smith. Practical lessons from Place Lab. *IEEE Pervasive Computing*, 5(3):32–39, 2006.

[14] MIT iFind website. http://ifind.mit.edu/, 2009.

[15] Ekahau Inc. Ekahau website. http://www.ekahau.com/, 2008.

[16] Skyhook Wireless Inc. Skyhook Wireless website. http://skyhookwireless.com/developers/, 2009.

[17] Spotigo Inc. Spotigo WiFi-based positioning solution. http://www.spotigo.com/products-and-services/spotigo-wifi-based-positioning-solution/, 2009.

[18] Yiming Ji, Sad Biaz, Santosh Pandey, and Prathima Agrawal. ARIADNE: a dynamic indoor signal map construction and localization system. *Proceedings of the Fourth International Conference on Mobile Systems, Applications, and Services (MobiSys 2006)*, pages 151–164, June 2006.

[19] Mikkel Baun Kjaergaard. A taxonomy for radio location fingerprinting. In *Location- and Context-Awareness*, volume 4718 of *Lecture Notes in Computer Science*, pages 139–156. Springer, Berlin / Heidelberg, 2007.

[20] John Krumm and John Platt. Minimizing calibration effort for an indoor 802.11 device location measurement system. Microsoft Research Technical Report MSR-TR-2003-82, November 2003.

[21] Nokia Beta Labs. Nokia activity monitor website. http://research.nokia.com/projects/activity_monitor, 2008.

[22] Andrew M. Ladd, Kostas E. Bekris, Guillaume Marceau, Lydia E. Kavraki, , and Dan S. Wallach. Robotics-based location sensing using wireless ethernet. *Proceedings of the 8th ACM International Conference on Mobile Computing and Networking (MOBICOM'02)*, pages 227–238, September 2002.

[23] Andrew M. Ladd, Kostas E. Bekris, Guillaume Marceau, Algis Rudys, Dan S. Wallach, and Lydia E. Kavraki. Using wireless ethernet for localization. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, 1:402–408, September 2002.

[24] Andrew M. Ladd, Kostas E. Bekris, Algis Rudys, Lydia E. Kavraki, , and Dan S. Wallach. Robotics-based location sensing using wireless ethernet. *Wireless Networks*, 11(1):189–204, January 2005.

[25] Andrew M. Ladd, Kostas E. Bekris, Algis Rudys, Dan S. Wallach, , and Lydia E. Kavraki. On the feasibility of using wireless ethernet for indoor localization. *IEEE Transactions on Robotics and Automation*, 20(3):555–559, June 2004.

[26] Andrew M. Ladd, Andreas Haeberlen, Eliot Flannery, Dan S. Wallach, Lydia E. Kavraki, and Algis Rudys. Practical robust localization over large-scale 802.11 wireless networks. *Proceedings of the 10th ACM International Conference on Mobile Computing and Networking (MOBICOM'04)*, pages 70–84, September 2004.

[27] Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, Fred Potter, Jason Tabert, Pauline Powledge, Gaetano Borriello, and Bill Schilit. Place Lab: Device positioning using radio beacons in the wild. *Proceedings of Pervasive 2005*, pages 116–133, 2005.

[28] Anthony LaMarca, Jeffrey Hightower, Ian Smith, and Sunny Consolvo. Self-mapping in 802.11 location systems. *Proceedings of Ubicomp 2005*, pages 87–104, September 2005.

[29] HyungJune Lee, Martin Wicke, and Branislav Kusy. Localization of mobile users using trajectory matching. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[30] Hendrik Lemelson, Thomas King, and Wolfgang Effelsberg. Pre-processing of fingerprints to improve the positioning accuracy of 802.11-based positioning systems. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[31] Julie Letchner, Dieter Fox, and Anthony LaMarca. Large-scale localization from wireless signal strength. *Proceedings of the National Conference on Artificial Intelligence (AAAI 2005)*, pages 15–20, 2005.

[32] Sense Networks. Sense Networks website. `http://www.sensenetworks.com/`, 2009.

[33] Andrew Ofstad, Emmett Nicholas, Rick Szcodronski, and Romit Roy Choudhury. AAMPL: Accelerometer augmented mobile phone localization. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[34] Hiranmay Parashar, Mark Thompson, and Ankur Bhattacharjee. Indoor localization over IEEE 802.11 using the Nokia N800. Software documentation at `http://themarkproject.com/wifilocator/`, May 2008.

[35] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The Cricket location-support system. *Proceedings of the 6th ACM International Conference on Mobile Computing and Networking (MOBICOM'00)*, pages 32–43, August 2000.

[36] Bill Schilit. Bootstrapping location-enhanced web services. Presentation at the University of Washington Computer Science Colloquia, December 2003.

[37] Nattapong Swangmuang and Prashant Krishnamurthy. On clustering RSS fingerprints for improving scalability of performance prediction of indoor positioning systems. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[38] Ping Tao, Algis Rudys, Andrew M. Ladd, , and Dan S. Wallach. Wireless LAN location-sensing for security applications. *Proceedings of the ACM Workshop on Wireless Security (WiSe)*, pages 11–20, September 2003.

[39] Seth Teller, Jonathan Battat, Ben Charrow, Dorothy Curtis, Russell Ryan, Jonathan Ledlie, and Jamey Hicks. Organic indoor location discovery. 2009. Submitted for publication in Proceedings of the Seventh International Conference on Mobile Systems, Applications, and Services (MobiSys 2009).

[40] Roy Want, Andy Hopper, Veronica Falco, and Jonathan Gibbons. The active badge location system. *ACM Transactions on Information Systems (TOIS)*, 10:91–102, January 1992.

[41] Eric W. Weisstein. Student's t-distribution. `http://mathworld.wolfram.com/Studentst-Distribution.html`, 2009. From MathWorld–A Wolfram Web Resource.

[42] Steve Whittaker, Loren G. Terveen, William C. Hill, and Lynn Cherny. The dynamics of mass interaction. In *Conference on Computer-Supported Cooperative Work (CSCW)*, pages 257–264, Seattle, WA, November 1998.

[43] Widyawan, Martin Klepal, Stéphane Beauregard, and Dirk Pesch. A novel backtracking particle filter for pattern matching indoor localization. Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT '08), September 2008.

[44] Wireless Geographic Logging Engine website. `http://www.wigle.net/`, 2009.

[45] Official WikiMapia website. `http://www.wikimapia.org/faq.htm`, 2009.

[46] Moustafa Youssef, Mohamed Abdallah, , and Ashok Agrawala. Multivariate analysis for WLAN location determination systems. *The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous 2005)*, pages 353–362, July 2005.

[47] Moustafa Youssef and Ashok Agrawala. Continuous space estimation for WLAN location determination systems. *IEEE Thirteenth International Conference on Computer Communications and Networks*, pages 161–166, October 2004.

[48] Moustafa Youssef and Ashok Agrawala. On the optimality of WLAN location determination systems. *Communication Networks and Distributed Systems Modeling and Simulation Conference*, January 2004.

[49] Moustafa Youssef and Ashok Agrawala. The Horus WLAN location determination system. *Proceedings of the Third International Conference on Mobile Systems, Applications, and Services (MobiSys 2005)*, pages 205–218, June 2005.

[50] Moustafa Youssef and Ashok Agrawala. Location-clustering techniques for energy-efficient WLAN location determination systems. *International Journal of Computers and Applications, 2005*, 2005.

[51] Moustafa Youssef and Ashok Agrawala. Analysis of the optimal strategy for WLAN location determination systems. *International Journal of Modelling and Simulation*, 27(1), 2007.

[52] Moustafa Youssef, Ashok Agrawala, and A. Udaya Shankar. WLAN location determination via clustering and probability distributions. *IEEE International Conference on Pervasive Computing and Communications (PerCom) 2003*, pages 143–150, March 2003.

[53] Moustafa Youssef, Adel Youssef, Chuck Rieger, Udaya Shankar, , and Ashok Agrawala. PinPoint: An asynchronous time-based location determination system. *Proceedings of the Fourth International Conference on Mobile Systems, Applications, and Services (MobiSys 2006)*, pages 165–176, June 2006.